

DOCUMENT RESUME

ED 429 146

UD 032 867

AUTHOR                   McMillan, James H.  
TITLE                   Establishing High Quality Classroom Assessments.  
INSTITUTION           Metropolitan Educational Research Consortium, Richmond, VA.  
PUB DATE              1999-03-00  
NOTE                   15p.  
PUB TYPE              Reports - Descriptive (141)  
EDRS PRICE           MF01/PC01 Plus Postage.  
DESCRIPTORS           Check Lists; \*Educational Assessment; Elementary Secondary Education; \*Scoring; \*Teacher Made Tests; Test Bias; \*Test Construction; Test Reliability; Test Validity

ABSTRACT

Ideas about educational assessment are synthesized to construct a Quality Assessment Checklist that can be used to evaluate the strengths and weaknesses of classroom assessments. Sound assessment begins with clear and appropriate learning targets. Once learning targets have been identified, they need to be matched with appropriate methods. Validity is another consideration, one that refers to the appropriateness of the inferences, uses, and consequences that result from the assessment. In classroom assessment, validity is determined by professional judgment. Reliability is concerned with the consistency, stability, and dependability of the scores. There is a number of things a teacher can do to enhance reliability in classroom assessment, including keeping procedures and scoring as objective as possible. Assessments must also be fair, giving all students an equal opportunity to demonstrate achievement. The consequences of the assessment must be considered, and the assessment must be designed to be practical and efficient. (SLD)

\*\*\*\*\*  
\*               Reproductions supplied by EDRS are the best that can be made       \*  
\*               from the original document.    \*  
\*\*\*\*\*

PERMISSION TO REPRODUCE AND  
DISSEMINATE THIS MATERIAL HAS  
BEEN GRANTED BY

G. Hupp  
MERC

TO THE EDUCATIONAL RESOURCES  
INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION  
Office of Educational Research and Improvement  
EDUCATIONAL RESOURCES INFORMATION  
CENTER (ERIC)

This document has been reproduced as  
received from the person or organization  
originating it.

Minor changes have been made to  
improve reproduction quality.

Points of view or opinions stated in this  
document do not necessarily represent  
official OERI position or policy.

# Establishing High Quality Classroom Assessments

By

James H. McMillan

Virginia Commonwealth University

March, 1999

BEST COPY AVAILABLE

## ESTABLISHING HIGH QUALITY CLASSROOM ASSESSMENTS

Until recently, the quality of classroom assessment was determined by the extent to which specific psychometric standards of validity, reliability, and efficiency were met. These standards were originally derived for large-scale, published, standardized objective tests, and are still important, at least conceptually, for most types of assessments. But in most classrooms such technical qualities have little relevance because the purpose of the assessment is different. This is not to say that the *ideas* of technical concepts such as validity and reliability are not important criteria for classroom assessment. High quality classroom assessment involves many other criteria as well, such as concerns about how the assessments influence student motivation and learning. For teachers, the focus is on the use and consequences of the results and what the assessments get students to do, rather than on a detailed inspection of the test itself.

High quality classroom assessments, then, are technically sound, provide results that demonstrate and improve targeted student learning, and inform instructional decision-making. Specific criteria can be derived from the Standards for Educational and Psychological Testing (1985), Standards for Teacher Competence in Educational Assessment of Students (1990), and recent work by a number of authors who have applied these more technical criteria to the realities of the classroom and needs of teachers (McMillan, 1997; Popham, 1995; Stiggins, 1997; Wiggins, 1997). What is summarized here is a synthesis of these ideas to construct a Quality Assessment Checklist that can be used to evaluate the strengths and weaknesses of classroom assessments (Figure 1).

Figure 1 here

### **Clear and Appropriate Learning Targets**

Sound assessment begins with clear and appropriate learning targets. It is best if these targets include both what students know and can do, and the criteria for judging student performance.

For example, students may be required to "understand how photosynthesis occurs." While this is a student outcome, much more is communicated by indicating something about the criteria for demonstrating understanding as well. Will they need to write a short essay about photosynthesis or answer multiple choice questions? By indicating both the knowledge and the criteria the specificity, clarity, and expectation of what is demonstrated is improved

Appropriateness of targets also refers to the effect of the target on students and how targets relate to more encompassing goals. Are the targets at the right level of difficulty to motivate students? Is there adequate balance with different types of targets? Are the targets consistent with teacher overall goals and the goals of the school and district? Are the targets comprehensive, covering all major dimensions that you hope to change and need feedback about?

### **Appropriateness of Assessment Methods**

Teachers have a number of different types of assessment methods at their disposal. Choosing the right ones depends on the strengths and weaknesses of the methods as well as the nature of the learning targets. That is, particular methods are more likely to provide quality assessments for certain types of targets. Once learning targets have been identified, they need to be matched with methods.

Figure 2 shows a matrix that summarizes the relative strengths of different methods in measuring different targets. Notice that for some targets several methods may be used. This is good in that it provides more flexibility in the assessments you use, but it also means there is no simple formula or "correct" method.

#### **Figure 2 here**

This matrix has been prepared to provide *general* guidelines about how well particular assessment methods measure each type of target. Variations to what is presented in the figure should be expected.

## Validity

Validity refers to the appropriateness of the inferences, uses, and consequences that result from the assessment. In other words, is the interpretation that is made from a test reasonable? Is the information gathered the right kind of evidence for the decision I need to make or the intended use? How sound is the interpretation of the information? Validity has to do with the consequences of the inferences, not the assessment itself. Thus, it is an inference or use that is valid or invalid, not the test, instrument, or procedure that is used to gather information. Often we use the phrase, "validity of the test," but it is more accurate to say "the validity of the interpretation, inference, or use of the results."

Validity in classroom assessment is determined by professional judgment. An analysis is done by accumulating evidence that would suggest that an inference or use is appropriate, and whether the consequences of the interpretations and uses are reasonable and fair. The most important evidence teachers draw upon to make inferences is based on the content of the assessment. Because it would not be feasible to assess students on everything they have learned, teachers typically select a *sample* of what has been taught to assess, and use student achievement on this sample to make inferences about knowledge of the entire universe or domain of content, reasoning, and other objectives. If the sample is judged to be representative of the universe or domain, then evidence based on content for validity is demonstrated. The inference from the test is that the student demonstrates knowledge about the unit.

Adequate sampling of content, determined by professional judgment, can be haphazard or very systematic. If there is a superficial review of the target, objectives, and test, validity is based only on *appearance*. (In the past this was sometimes referred to as face validity.) Once the complete domain of content and objectives is specified, the items on the test can be reviewed to be certain that there is a match between the intended inferences and what is on the test.

Another consideration is the extent to which an assessment can be said to have *instructional*

validity. Instructional validity is concerned with the match between what is taught and what is assessed. How closely does the test correspond to what has been covered in class and in assignments? Have students had the opportunity to learn what has been assessed? This type of evidence for validity is important for making reasonable inferences about student performance.

### **Reliability**

Reliability is concerned with the consistency, stability, and dependability of the scores. While it is generally understood that classroom assessments contain some degree of error, it is rare that teachers obtain a specific indication of the amount of error. Nevertheless, the idea that there is error in assessment is critical. It is not a matter of "all or none," as if some results are reliable and others unreliable. Rather, for each assessment there is some *degree* of error. Thus, we think in terms of low, moderate, or high reliability. Reliability is influenced by factors within the student (internal sources of error), such as luck, mood, and physical condition, as well as external factors such as the quality of the test, sampling of items, scoring errors, and test directions. The direction of the error can be positive or negative. Sometimes it is obvious when a student's score is lower than it should be based on the behavior of the student at the time of the assessment, e.g., if the student was sick, tired, in a bad mood, or distracted. Of particular concern with many performance-based assessments is error that occurs due to observer bias. When teachers evaluate through observation there are a number of factors that can contribute to error, such as observer fatigue and expectations.

There are a number of assessment procedures that teachers can use to enhance reliability and reduce error. First, there needs to be a sufficient number of items or tasks. The greater the number of items and the number of times a trait is assessed, the more reliable the inference about what a student knows and can do. Practically speaking, this means that teachers should rarely, if ever, rely on a single assessment. However the opposite is not necessarily true. That is, teachers should not give as many assessments as possible! A reasonable rule of thumb is to have 8-10

items measure a single trait and three or four major assessments during a unit. Avoid having a single assessment count too much, since any error with that assessment is then also counted heavily. Assessment should continue until a consistent pattern of results is obtained, using many shorter assessments rather than fewer long assessments. Second, keep assessment procedures and scoring as objective as possible to avoid bias. Obviously "machine scorable" tests are more objective with respect to scoring (though subjectivity is needed to write items). For performance-based assessments this means having a clear rubric. Third, do everything possible to eliminate the influence of extraneous sources of error. Thus, interruptions should be avoided, temperature and lighting should be appropriate, directions should be clear, and so forth. Finally, it is best to use different assessment methods since each method has unique sources of error and students may be more adept with some methods and struggle with others. This suggests that teachers should not rely solely on one method, such as giving all multiple choice exams or all essays.

### Fairness

A fair assessment is one that provides an equal opportunity to all students to demonstrate achievement. We want to allow students to show us what they have learned from instruction. If some students have an advantage over others because of factors unrelated to what is being taught, then the assessment is not fair. Fair assessments are *unbiased* and *nondiscriminatory*. That is, fairness means that neither the assessment task nor scoring is differentially affected by race, gender, ethnic background, handicapping condition, or other factors unrelated to what is being assessed. Fairness is also evident in what students are told about the assessment and whether students have had the opportunity to learn what is being assessed.

A fair assessment is one in which it is clear what will and will not be tested. Both the content of the assessment and the scoring criteria should be *public*. Being public means that students know the content and scoring criteria prior to the assessment and often prior to instruction.

When students know what will be assessed they know what to study and focus on. By knowing the scoring criteria students understand much better the qualitative differences the teacher is looking for in student performance. One way to help students understand the assessment is to give them the assessment blueprint, sample questions, and examples of work completed by previous students and graded by the teacher.

Is it fair to assess students on things they have not had the opportunity to learn? Opportunity to learn means that students know what to learn and then are provided ample time and appropriate instruction. It is usually not sufficient to simply tell students what will be assessed and then test them. Instruction must be planned that focuses specifically on helping students understand, providing feedback to students on their progress, and providing the time needed to learn.

It is unfair to assess students on things that require prerequisite knowledge or skills that they do not possess? This means that teachers need to have a good understanding of needed prerequisites. For example, suppose a teacher is testing math reasoning skills. The questions are based on short paragraphs that provide needed information. In this situation math reasoning skills can be demonstrated only if students can read and understand the paragraphs. Thus, reading skills are prerequisites. If students do poorly on the assessment their performance may have more to do with a lack of reading skills than with math reasoning.

Another type of prerequisite skill is concerned with test-taking. Some students bring better test-taking skills to an assessment than other students, such as knowing to read directions carefully, pacing, initially bypassing difficult items, checking answers, and eliminating wrong answers to multiple choice items rather than looking for the right answer. These skills are not difficult for students to learn, and it is advisable to make sure all students are familiar with these skills prior to assessment.

Another source of bias can be found in the nature of the actual assessment task - the contents and process of the test, project, problem, or other task. Bias is present if the assessment distorts

performance due to the student's ethnicity, gender, race, religious background, and so on.

Popham (1995) has identified two major forms of assessment bias - *offensiveness* and *unfair penalization*. Offensiveness occurs if the content of the assessment offends, upsets, distresses, angers, or otherwise creates negative affect for particular students or a subgroup of students. Offensiveness occurs most often when stereotypes of particular groups are present in the assessment. Suppose a test question portrayed a minority group in low-paying, low status jobs, and anglo groups in high-paying, high status jobs. Students who are members of the minority group may understandably be offended by the question, mitigating their performance

Unfair penalization is bias that disadvantages a student because of content that makes it more difficult for students from some groups to perform as compared to students from other groups. That is, bias is evident when an unfair advantage or disadvantage is given to one group because of gender, socioeconomic status, race, language, or other characteristic. Suppose students take an "aptitude" test that uses rural, farm-oriented examples. The questions deal with types of cows and pigs, winter wheat, and farm equipment. If the student grew up in a suburban environment, will his or her score be as good as students who grew up on a farm?

Cultural differences that are reflected in vocabulary, prior experiences, skills, and values may influence the assessment. These differences are especially important in our increasingly diverse society and classroom. For example, depending on the culture, rules for sharing beliefs, discussion, taking turns, and expressing opinions differ. Respect and politeness may be expressed differently by students from different backgrounds (e.g., not looking into your eyes, silence, squinting as a way to say "no", looking up or down when asked a question).

Another type of assessment task bias that has received a lot of attention recently is the need to accommodate the special abilities of exceptional children. An assessment is biased if performance is affected by a disability or other limiting characteristic when the student actually possesses the knowledge or skill being measured. In other words, when assessing exceptional students teachers need to modify the assessment task so that the handicapping trait is not a factor

in the performance. For example, students with a hearing deficiency may need written directions for completing an assessment that you give orally to other students.

### **Positive Consequences**

The nature of classroom assessments has important consequences for teaching and learning. The most direct consequence of assessment is that students learn and study in a way that is consistent with your assessment task. If the assessment is a multiple choice test of knowledge of specific facts, then students will tend to memorize information. On the other hand, if the assessment calls for extended essays students tend to learn the material in larger, related "chunks," and practice recall rather than recognition when studying. Assessments that require problem-solving, such as performance-based assessments, encourage students to think and apply what they learn.

Assessments also have clear consequences on student motivation. If students know what will be assessed, how it will be scored, and believe that it will be fair, they are likely to be more motivated to learn. Motivation is also increased when the assessment tasks are relevant to student backgrounds and goals, challenging but possible, and structured to give them individualized feedback about their performance. What good is a high score on an easy test? Authentic assessments typically provide more active learning, which increases motivation. Giving students multiple assessments, rather than a single assessment, lessens fear and anxiety. With less apprehension, risk-taking, exploration, creativity, and questioning are enhanced.

The student-teacher relationship is influenced by the nature of assessment. When teachers construct assessments carefully and provide feedback to students, the relationship is strengthened. Conversely, if students have the impression that the assessment is sloppy, not matched with course objectives, designed to trick students (like some true/false questions we have all answered!), and provide little feedback, the relationship is weakened. Assessment affects the way students perceive the teacher and gives them an indication of how much the teacher

cares about them and what they learn.

Like students, teachers are affected by the nature of the assessments they give their students. Just as students learn depending on the assessment, teachers tend to teach to the test (not teach *the* test, teach *to* the test). Thus, if the assessment calls for memorization of facts, the teacher tends to teach lots of facts; if the assessment requires reasoning, then the teacher structures exercises and experiences that get students to think. Teachers need to know about the assessment methods they select. This includes knowledge of the strengths and limitations of the method, how to administer the assessment, how to score and properly interpret student responses, the time needed to construct the assessment and score the results, and appropriateness of the method for given learning targets.

### **Practicality and Efficiency**

High quality assessments are practical and efficient. Time is a limited commodity for teachers, and the pace of classrooms is hectic. Assessments should not detract from energy that is needed for instruction and other professional activities. For example, while it may seem best to use extensive performance-based assessments for reasons associated with validity and fairness, these assessments may take too much teacher time to prepare, observe, and score. Factors that effect practicality and efficiency include teacher familiarity with the method, time required, complexity of administration, and ease of scoring and interpretation.

Teachers need to be knowledgeable and skilled in the assessments they use. This suggests a need for teacher self-evaluation, monitoring by others, and training and opportunities that allow teachers to brush up on their assessment skills. All too often teachers are simply expected to be able to administer most any kind of assessment without adequate training (many teachers never even take a classroom assessment course).

Assessments that take a long time in set up and have long, complicated directions are less efficient than other types of assessments that require simple administration, such as objective

tests. Some assessments are also easier to score and interpret. For subjectively scored assessments it is efficient to use a key that can communicate different levels of performance and provide meaningful feedback.

### **Summary**

High quality classroom assessments provide reliable, valid, and useful measures of student performance. Quality is enhanced when the assessments balance technical requirements with practical consequences. As assessment becomes more important the quality of these assessments becomes crucial. Teachers need to know what constitutes high quality assessment and to be judged against these criteria. The result will be enhanced professionalism, increased student learning, and more credible reports of student learning.

## References

- McMillan, James H. 1997. *Classroom assessment: Principles and practice for effective instruction.* Needham Heights, MA: Allyn & Bacon.
- Popham, W. James. 1995. *Classroom assessment: What teachers need to know,* Needham Heights, MA: Allyn & Bacon.
- . *Standards for educational and psychological testing,* 1985. Washington, DC: American Psychological Association.
- . *Standards for teacher competence in educational assessment of students,* 1990. American Federation of Teachers, National Council on Measurement in Education, National Education Association.,
- Stiggins, Richard J. 1997. *Student-centered classroom assessment,* 2<sup>nd</sup> ed. Upper Saddle River, NJ: Prentice-Hall, Inc..
- Wiggins, Grant. 1997. Practicing what we preach in designing authentic assessments, *Educational Leadership*, 54, 18-25.

- ✓ Clear and Appropriate Learning Targets
- ✓ Appropriateness of Assessment Methods
- ✓ Validity
- ✓ Reliability
- ✓ Fairness
- ✓ Positive Consequences
- ✓ Practicality and Efficiency

Figure 1: Quality Assessment Checklist.

<b>Targets</b>	<b>Assessment Methods</b>					
	<u>Objective</u>	<u>Essay</u>	<u>Performance-Based</u>	<u>Oral Question</u>	<u>Observation</u>	<u>Self report</u>
Knowledge	5	4	3	4	3	2
Reasoning	2	5	4	4	2	2
Skills	1	3	5	2	5	3
Products	1	1	5	2	4	4

Figure 2. Matching Learning Targets With Methods of Assessment.<sup>1</sup>

note: Higher numbers indicate better matches.

---

<sup>1</sup> Adapted from McMillan, 1997, p.50.



**U.S. Department of Education**  
**Office of Educational Research and Improvement (OERI)**  
**National Library of Education (NLE)**  
**Educational Resources Information Center (ERIC)**



## REPRODUCTION RELEASE

(Specific Document)

### I. DOCUMENT IDENTIFICATION:

Title: Establishing High Quality Classroom Assessment

Author(s): James H. McMillan

Corporate Source:

Virginia Commonwealth University

Publication Date:

June 1998

### II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract Journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of this page.

The sample sticker shown below will be attached to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY	
<i>Sample</i>	
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	
1	

Level 1

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be attached to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN OTHER ERICONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY	
<i>Sample</i>	
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	
2A	

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be attached to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY	
<i>Sample</i>	
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	
2B	

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.

If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, →  
base

Signature: *James H. McMillan*

Printed Name/Position/Title:

James McMillan / Professor

Organization/Address: Box 842020, Virginia Commonwealth U., Richmond VA 23284-2020

Telephone: 804 8281332

FAX: 804 82253554

E-mail Address: *jmmill@vcu.edu*

Date: 8/24/99

Date: Mon, 23 Aug 99 8:46:38 EDT  
To: <crowsome@inet.ed.gov>  
From: "James H. McMillan" <jmcmillan@edunet.soe.vcu.edu>  
Reply-To: <jmcmillan@edunet.soe.vcu.edu>  
Subject: re: Your Paper Submitted to ERIC  
X-Incognito-SN: 1728  
X-Incognito-Version: 4.11.23

This is not an official MERC paper, so I will need to sign a release form. Please fax me a copy at 804 225 3554. Thanks.

----- Original Text -----

From: "Christina Rowsome" <crowsome@inet.ed.gov>, on 8/19/99 2:35 PM:

Dr. McMillan:

A paper that you authored, "Establishing High Quality Classroom Assessments" is being processed for the ERIC database. If it is an official paper of the Metropolitan Educational Research Consortium at VCU it can be processed under the ERIC Reproduction Release that ERIC has with MERC. If it is not a MERC publication, we need an ERIC Reproduction Release signed by you as the individual contributor. Please advise me of the status of this paper. (ERIC Reference # UD032867)

Thanks!